Business Research Methods: Data Analysis- II

By Dr. Satyabrata Dash

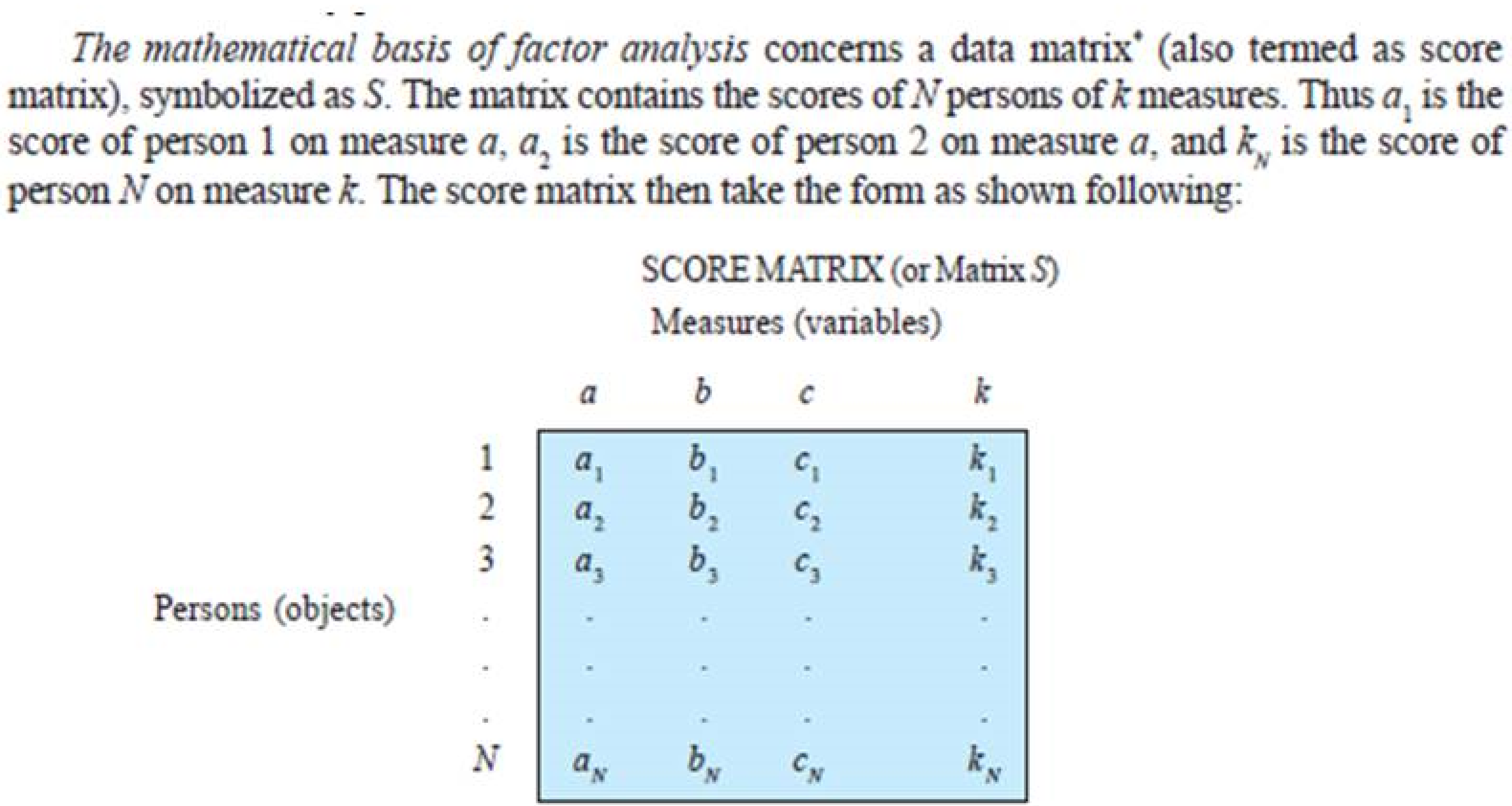
Professor- MBA Marketing

SMIT- PGCMS, Brahmapur

Factor analysis

- Factor analysis is by far the most often used multivariate technique of research studies, specially pertaining to social and behavioral sciences. It is a technique applicable when there is a systematic interdependence among a set of observed or manifest variables and the researcher is interested in finding out something more fundamental or latent which creates this commonality.
- For instance, we might have data, say, about an individual's income, education, occupation and dwelling area and want to infer from these some factor (such as social class) which summarizes the commonality of all the said four variables.
- The technique used for such purpose is generally described as factor analysis. Factor analysis, thus, seeks to resolve a large set of measured variables in terms of relatively few categories, known as factors.
- This technique allows the researcher to group variables into factors (based on correlation between variables) and the factors so derived may be treated as new variables (often termed as latent variables) and their value derived by summing the values of the original variables which have been grouped into the factor. The meaning and name of such new variable is subjectively determined by the researcher. Since the factors happen to be linear combinations of data, the coordinates of each observation or variable is measured to obtain what are called factor loadings. Such factor loadings represent the correlation between the particular variable and the factor, and are usually place in a matrix of correlations between the variable and the factors.

Factor analysis



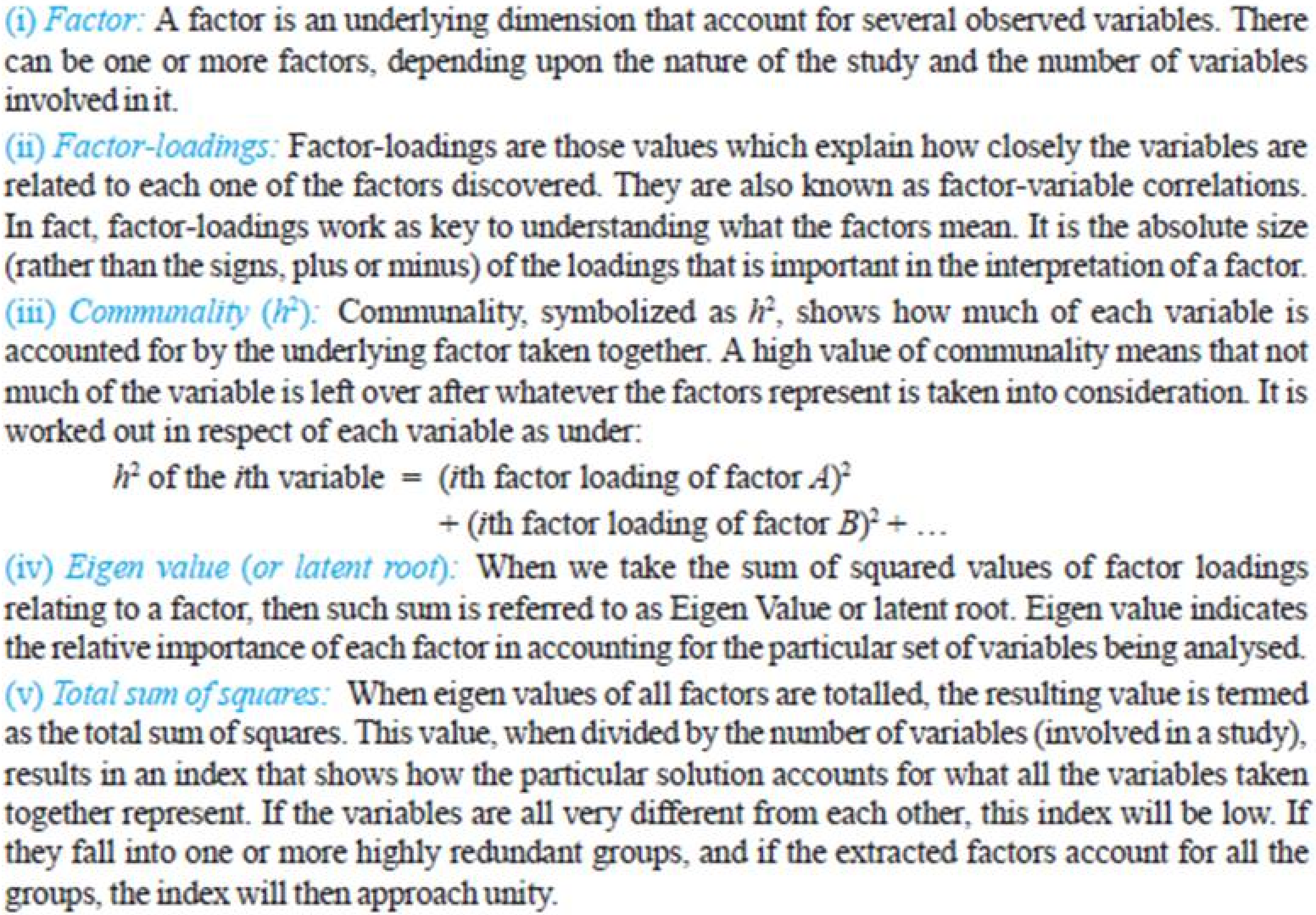
The mathematical basis of factor analysis concerns a data matrix* (also termed as score matrix), symbolized as S. The matrix contains the scores of N persons of k measures. Thus a_1 is the score of person 1 on measure a, a_2 is the score of person 2 on measure a, and k_N is the score of person N on measure k. The score matrix then take the form as shown following:

SCORE MATRIX (or Matrix S)

Measures (variables)

$$\begin{array}{llccccc} & & a & b & c & & k \\ & 1 & a_1 & b_1 & c_1 & & k_1 \\ & 2 & a_2 & b_2 & c_2 & & k_2 \\ & 3 & a_3 & b_3 & c_3 & & k_3 \\ \text{Persons (objects)} & . & . & . & . & & . \\ & . & . & . & . & & . \\ & . & . & . & . & & . \\ & N & a_N & b_N & c_N & & k_N \end{array}$$

Terms used in Factor analysis



(i) *Factor:* A factor is an underlying dimension that account for several observed variables. There can be one or more factors, depending upon the nature of the study and the number of variables involved in it.

(ii) *Factor-loadings:* Factor-loadings are those values which explain how closely the variables are related to each one of the factors discovered. They are also known as factor-variable correlations. In fact, factor-loadings work as key to understanding what the factors mean. It is the absolute size (rather than the signs, plus or minus) of the loadings that is important in the interpretation of a factor.

(iii) *Communality* (h^2): Communality, symbolized as h^2, shows how much of each variable is accounted for by the underlying factor taken together. A high value of communality means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under:

$$h^2 \text{ of the } i\text{th variable} = (i\text{th factor loading of factor } A)^2 + (i\text{th factor loading of factor } B)^2 + \ldots$$

(iv) *Eigen value (or latent root):* When we take the sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen Value or latent root. Eigen value indicates the relative importance of each factor in accounting for the particular set of variables being analysed.

(v) *Total sum of squares:* When eigen values of all factors are totalled, the resulting value is termed as the total sum of squares. This value, when divided by the number of variables (involved in a study), results in an index that shows how the particular solution accounts for what all the variables taken together represent. If the variables are all very different from each other, this index will be low. If they fall into one or more highly redundant groups, and if the extracted factors account for all the groups, the index will then approach unity.

Centroid Method of Factor Analysis (Example)

R		Variables (Measures)							
		1	2	3	4	5	6	7	8
Variables (Group of Persons)	1	1.000	0.709	0.204	0.081	0.626	0.113	0.155	0.774
	2	0.709	1.000	0.051	0.089	0.581	0.098	0.083	0.652
	3	0.204	0.051	1.000	0.671	0.123	0.689	0.582	0.072
	4	0.081	0.089	0.671	1.000	0.022	0.798	0.613	0.111
	5	0.626	0.581	0.123	0.022	1.000	0.047	0.201	0.724
	6	0.113	0.098	0.689	0.798	0.047	1.000	0.801	0.120
	7	0.155	0.083	0.582	0.613	0.201	0.801	1.000	0.152
	8	0.774	0.652	0.072	0.111	0.724	0.120	0.152	1.000

R		Variables (Measures)							
		1	2	3	4	5	6	7	8
Variables (Persons)	1	1.000	0.709	0.204	0.081	0.626	0.113	0.155	0.774
	2	0.709	1.000	0.051	0.089	0.581	0.098	0.083	0.652
	3	0.204	0.051	1.000	0.671	0.123	0.689	0.582	0.072
	4	0.081	0.089	0.671	1.000	0.022	0.798	0.613	0.111
	5	0.626	0.581	0.123	0.022	1.000	0.047	0.201	0.724
	6	0.113	0.098	0.689	0.798	0.047	1.000	0.801	0.120
	7	0.155	0.083	0.582	0.613	0.201	0.801	1.000	0.152
	8	0.774	0.652	0.072	0.111	0.724	0.120	0.152	1.000
Column Sums =		3.662	3.263	3.392	3.385	3.324	3.666	3.587	3.605

Sum of Column sums (T) =	27.884
√T =	5.281

First Centroid factor A= Column Sum/ √T (given below)

1	2	3	4	5	6	7	8
0.693	0.618	0.642	0.641	0.629	0.694	0.679	0.683

Second Centroid factor B (Row x Column) each

Q1		1	2	3	4	5	6	7	8
		0.693	0.618	0.642	0.641	0.629	0.694	0.679	0.683
1	0.693	0.481	0.429	0.445	0.445	0.437	0.481	0.471	0.473
2	0.618	0.429	0.382	0.397	0.396	0.389	0.429	0.420	0.422
3	0.642	0.445	0.397	0.413	0.412	0.404	0.446	0.436	0.439
4	0.641	0.445	0.396	0.412	0.411	0.404	0.445	0.435	0.438
5	0.629	0.437	0.389	0.404	0.404	0.396	0.437	0.428	0.430
6	0.694	0.481	0.429	0.446	0.445	0.437	0.482	0.472	0.322
7	0.679	0.471	0.420	0.436	0.435	0.428	0.472	0.461	0.464
8	0.683	0.473	0.422	0.439	0.438	0.430	0.474	0.464	0.466

First Matrix of Residual Coefficient = (R - Q1)

R1	1	2	3	4	5	6	7	8
1	0.519	0.280	-0.241	-0.364	0.189	-0.368	-0.316	0.301
2	0.280	0.618	-0.346	-0.307	0.192	-0.331	-0.337	0.230
3	-0.241	-0.346	0.587	0.259	-0.281	0.243	0.146	-0.367
4	-0.364	-0.307	0.259	0.589	-0.382	0.353	0.178	-0.327
5	0.189	0.192	-0.281	-0.382	0.604	-0.390	-0.227	0.294
6	-0.368	-0.331	0.243	0.353	-0.390	0.518	0.329	-0.202
7	-0.316	-0.337	0.146	0.178	-0.227	0.329	0.539	-0.312
8	0.301	0.230	-0.367	-0.327	0.294	-0.354	-0.312	0.534

N.B

Reflecting the variables 3, 4, 6 and 7, we obtain reflected matrix of residual coefficient (R'1) as Reflecting the variables 3, 4, 6 and 7, we obtain reflected matrix of residual coefficient (R'1) as under

Reflected Matrix of Residual Coefficients (Ignoring - and make it + numbers)

R'1	1	2	3*	4*	5	6*	7*	8
1	0.519	0.280	0.241	0.364	0.189	0.368	0.316	0.301
2	0.280	0.618	0.346	0.307	0.192	0.331	0.337	0.230
3	0.241	0.346	0.587	0.259	0.281	0.243	0.146	0.367
4	0.364	0.307	0.259	0.589	0.382	0.353	0.178	0.327
5	0.189	0.192	0.281	0.382	0.604	0.390	0.227	0.294
6	0.368	0.331	0.243	0.353	0.390	0.518	0.329	0.202
7	0.316	0.337	0.146	0.178	0.227	0.329	0.539	0.312
8	0.301	0.230	0.367	0.327	0.294	0.354	0.312	0.534
Column Sums =	2.579	2.642	2.471	2.758	2.559	2.887	2.382	2.566

Sum of Column Sums (T) =	20.843
√T =	4.565

Second Centroid factor B= Column Sum/ √T (given below)

	1	2	3*	4*	5	6*	7*	8
	0.565	0.579	0.541	0.604	0.561	0.632	0.522	0.562
(-ve) to * mark columns	0.565	0.579	-0.541	-0.604	0.561	-0.632	-0.522	0.562

Now we can write the matrix of factor loadings as under:

	Factor Loadings	
Variables	Centroid Factor A	Centroid Factor B
1	0.693	0.565
2	0.618	0.579
3	0.642	-0.541
4	0.641	-0.604
5	0.629	0.561
6	0.694	-0.632
7	0.679	-0.522
8	0.683	0.562

Calculate Communality (h2) from factor loading matrix = (Centroid Factor Centroid Factor -A)^2 + (Centroid Factor Centroid Factor-B)^2

	Factor Loadings		
Variables	Centroid Factor A	Centroid Factor B	Communality (h2)
1	0.693	0.565	0.800
2	0.618	0.579	0.717
3	0.642	-0.541	0.705
4	0.641	-0.604	0.776
5	0.629	0.561	0.711
6	0.694	-0.632	0.881
7	0.679	-0.522	0.734
8	0.683	0.562	0.782
Communality (h2) =			6.106

Calculate Eigen value as common variance (ά) = Sum of Sqared values of factor loadings as below

	Factor Loadings			
Variables	(Centroid Factor A)^2	(Centroid Factor B)^2	Communality (h2)	
1	0.481	0.319	0.800	
2	0.382	0.335	0.717	
3	0.413	0.293	0.705	
4	0.411	0.365	0.776	
5	0.396	0.315	0.711	
6	0.482	0.399	0.881	
7	0.461	0.272	0.734	
8	0.466	0.316	0.782	
Eigen value (ά) =	3.492	2.614	6.106	NB: (ά1+ά2)= ∑h2
Proportion of Total variance =	0.44	0.33	0.76	NB: Eigen value (ά)/ N
	44%	33%	77%	
Proportion of Common variance =	0.57	0.43	1.0	Calculate on the basis of 1.0
	57%	43%	100%	

Interpretation

The total variance (V) in the analysis is taken as equal to the number of variables involved (on the presumption that variables are standardized). In this present example, then V = 8.0. The row labeled “Eigen value” or “Common variance” gives the numerical value of that portion of the variance attributed to the factor in the concerning column above it. These are found by summing up the squared values of the corresponding factor loadings. Thus the total value, 8.0, is partitioned into 3.492 as Eigen value for factor A and 2.614 as Eigen value for factor B and the total 6.106 as the sum of Eigen values for these two factors. The corresponding proportion of the total variance, 8.0, are shown in the next row; there we can notice that 77% of the 330 Research Methodology total variance is related to these two factors, i.e., approximately 77% of the total variance is common variance whereas remaining 23% of it is made up of portions unique to individual variables and the techniques used to measure them. The last row shows that of the common variance approximately 57% is accounted for by factor A and the other 43% by factor B. Thus it can be concluded that the two factors together “explain” the common variance.

Multiple Regression Analysis (R)

- Multiple regression analysis is a straightforward extension of simple regression analysis which allows more than one independent variable. This extension, however, makes multiple regression analysis an incredibly versatile tool that can be used in an enormous variety of statistical problems. A reasonably complete treatment of multiple regression would require a book at least as long as this one already is. This chapter will deal with some of the issues and techniques that frequently arise with multiple regression analysis. Multiple regression analysis provides the analyst with such a variety of techniques that the primary problem is to decide exactly what form the regression equation (or regression *model)* will take, including which independent variables will be used.

Qn.

Independent Variable1 (X1)	Independent Variable2 (X2)	Dependent Variable (Y)
Highest year of School completed	Motivation	Annual Sales (Rs)
12	28	350,400
13	35	399,765
15	45	429,000
11	55	435,000
18	65	433,200

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

Correlation between Highest Year of School and Motivation (rx1,x2) =

	Independent Variable1 (X1)	Independent Variable2 (X2)					
	Highest year of School completed	Motivation	x1= X1-MeanX1	x2= X2-MeanX2	$x1^2$	$x2^2$	x1x2
	12	28	-1.8	-17.6	3.24	309.76	31.68
	13	35	-0.8	-10.6	0.64	112.36	8.48
	15	45	1.2	-0.6	1.44	0.36	-0.72
	11	55	-2.8	9.4	7.84	88.36	-26.32
	18	65	4.2	19.4	17.64	376.36	81.48
Total=	69	228	0	0	30.8	887.2	94.6
n=	5	5					
Mean=	13.8	45.6		$\sum x1^2 * \sum x2^2$=	27325.76		
$\sqrt{\sum x1^2 * \sum x2^2}$=	165.305						
rx1x2=	0.572	$(rx1x2)^2$=	0.327				

Correlation between Highest Year of School and Motivation (rx2,y) =

	Independent Variable2 (X2)	Dependent Variable (Y)					
	Motivation	Annual Sales (Rs)	x2= X2-MeanX2	y= Y-MeanY	$x2^2$	y^2	x2y
	28	350400	-17.6	-59073	309.76	3.49E+09	1039685
	35	399765	-10.6	-9708	112.36	94245264	102904.8
	45	429000	-0.6	19527	0.36	3.81E+08	-11716.2
	55	435000	9.4	25527	88.36	6.52E+08	239953.8
	65	433200	19.4	23727	376.36	5.63E+08	460303.8
Total=	228	2047365	0	0	887.2	5.18E+09	1831131
n=	5	5					
Mean=	45.6	409473		$\sum x2^{2}*\sum y^{2}=$	4.6E+12		
$\sqrt{\sum x2^{2}*\sum y^{2}}=$	2143709.148						
rx2y=	0.854	$(rx2y)^2=$	0.730				

Correlation between Highest Year of School and Motivation (rx1,y) =

	Independent Variable1 (X1)	Dependent Variable (Y)					
	Highest year of School completed	Annual Sales (Rs)	x1= X1-MeanX1	y= Y-MeanY	$x1^2$	y^2	x1y
	12	350400	-1.8	-59073	3.24	3.49E+09	106331.4
	13	399765	-0.8	-9708	0.64	94245264	7766.4
	15	429000	1.2	19527	1.44	3.81E+08	23432.4
	11	435000	-2.8	25527	7.84	6.52E+08	-71475.6
	18	433200	4.2	23727	17.64	5.63E+08	99653.4
Total=	69	2047365	0	0	30.8	5.18E+09	165708
n=	5	5					
Mean=	13.8	409473		$\sum x1^{2}*\sum y^{2}=$	1.6E+11		
$\sqrt{\sum x1^{2}*\sum y^{2}}=$	399420.594						
rx1y=	0.415	$(rx1y)^2=$	0.172				

Calculate b1 and b2=

$$b_1 = \left(\frac{r_{y,x1} - r_{y,x2} r_{x1,x2}}{1 - \left(r_{x1,x2} \right)^2} \right) \left(\frac{SD_y}{SD_{x1}} \right)$$

$$b_2 = \left(\frac{r_{y,x2} - r_{y,x1} r_{x1,x2}}{1 - \left(r_{x1,x2} \right)^2} \right) \left(\frac{SD_y}{SD_{x2}} \right)$$

b1=	-1426.208058
b2=	0.035241388

SDx1 =	2.7748887385
SDx2 =	14.892295135
Sdy =	35985.29762
rx1x2=	0.572275313
rx2y=	0.854188173
rx1y=	0.414870947

Calculate a =

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

Mean X1	13.8
Mean X2	45.6
Mean Y	409473

a =	429153.0642

Calculate Y =

$$Y' = a + b_1X_1 + b_2X_2$$

X1 = Year of School =	13	Prediction of values
X2 = Motivation Score =	49	

Y =	410614.0863

So, given a job applicant with 13 years of education completed and who received a motivation score of 49 on the Higgins Motivation Scale, our single best prediction of how much this person will earn for our dealership is $685,881.74.

Discriminant Analysis

- Discriminant Analysis finds a set of prediction equations based on independent variables that are used to classify individuals into groups. There are two possible objectives in a discriminant analysis: finding a predictive equation for classifying new individuals or interpreting the predictive equation to better understand the relationships that may exist among the variables.
- Discriminant function analysis is used to classify individuals into the predetermined groups. It is a multivariate analogue of analysis of variance, and can be considered as an *a posterior procedure of multivariate analysis of variance*
- If discriminant function analysis is effective for a set of data, the classification table of correct and incorrect estimates will yield a high percentage correct.
- Multiple discriminant function analysis (sometimes called canonical variety analysis) is used when there are three or more groups.
- Discriminant analysis (in the broad sense) is a very powerful statistical tool for many types of analyses.

Technical Details

Suppose you have data for K groups, with N_k observations per group. Let N represent the total number of observations. Each observation consists of the measurements of p variables. The i^{th} observation is represented by X_{ki}. Let M represent the vector of means of these variables across all groups and M_k the vector of means of observations in the k^{th} group.

Define three sums of squares and cross products matrices, S_T, S_W, and S_A, as follows

$$S_T = \sum_{k=1}^{K}\sum_{i=1}^{N_k}(X_{ki} - M)(X_{ki} - M)'$$

$$S_W = \sum_{k=1}^{K}\sum_{i=1}^{N_k}(X_{ki} - M_k)(X_{ki} - M_k)'$$

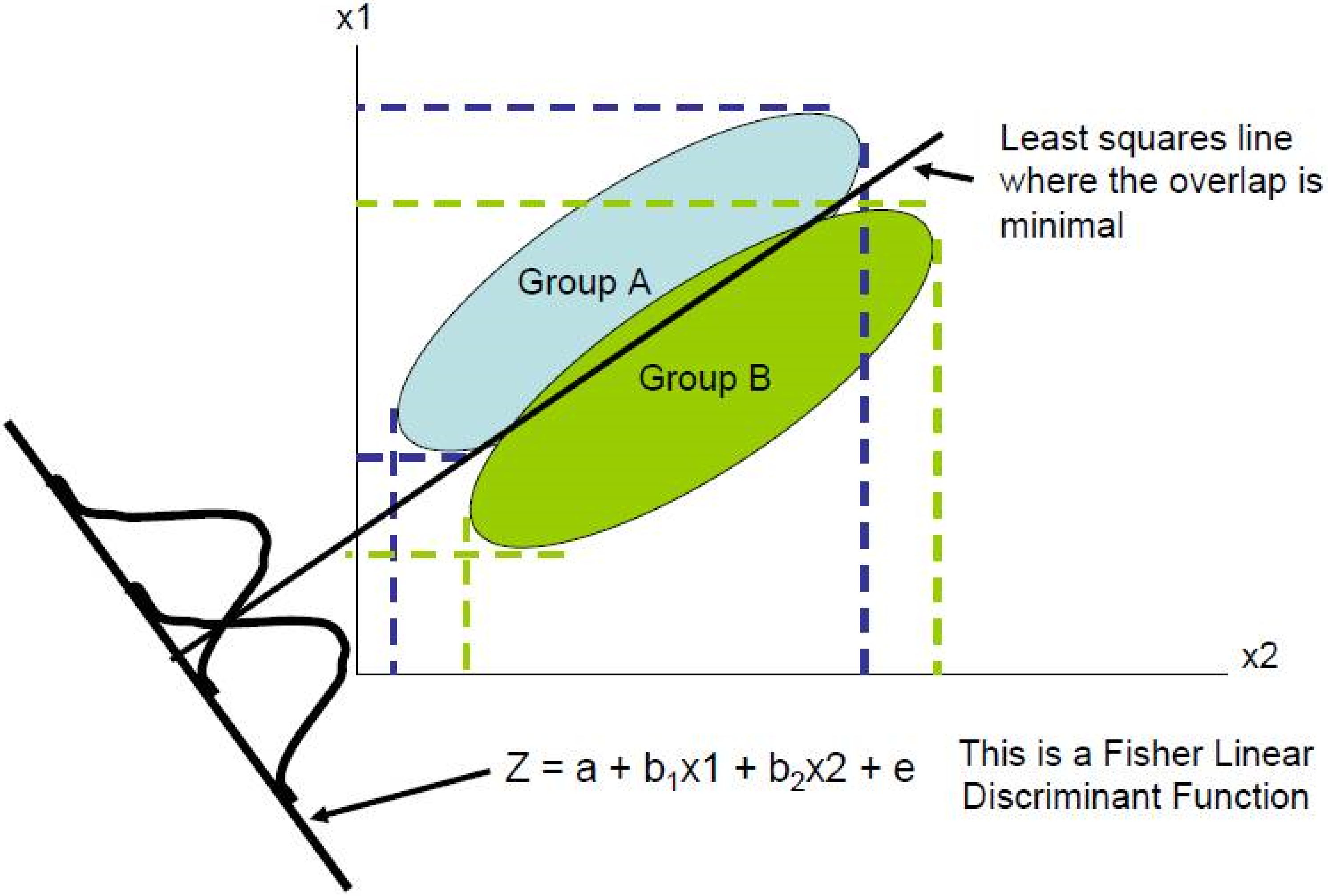
$$S_A = S_T - S_W$$

Next, define two degrees of freedom values, *df1* and *df2*:

$$df1 = K - 1$$

$$df2 = N - K$$

The Discriminant Analysis Problem in 2-Space



Least squares line where the overlap is minimal

$Z = a + b_1x1 + b_2x2 + e$

This is a Fisher Linear Discriminant Function